Partially-Hidden Markov Models

Emmanuel Ramasso and Thierry Denœux and Noureddine Zerhouni

Abstract This paper addresses the problem of Hidden Markov Models (HMM) training and inference when the training data are composed of feature vectors plus uncertain and imprecise labels. The "soft" labels represent partial knowledge about the possible states at each time step and the "softness" is encoded by belief functions. For the obtained model, called a Partially-Hidden Markov Model (PHMM), the training algorithm is based on the Evidential Expectation-Maximisation (E2M) algorithm. The usual HMM model is recovered when the belief functions are vacuous and the obtained model includes supervised, unsupervised and semi-supervised learning as special cases.

1 Introduction

Hidden Markov Models (HMM) are powerful tools for sequential data modelling and analysis. Many applications for several decades have found solutions based on HMM such as discovering word sequences based on speech audio recordings [9], gene finding based on a DNA sequence [8], and performing prognostics and health detection of ball bearings degradation based on noisy sensors [6, 10]. In the sequel, we consider sequential data taking the form of a time-series of length T where each element is a multidimensional feature vector $x_t \in \Re^F, t = 1 \dots T$ also called vector of observations [9]. The modelling part assumes that the system (a speaker, a DNA

E. Ramasso and N. Zerhouni
FEMTO-ST Institute, UMR CNRS 6174 - UFC / ENSMM / UTBM, Automatic Control and Micro-Mechatronic Systems Department, 24 rue Alain Savary, F-25000 Besançon, France, e-mail: emmanuel.ramasso@femto-st.fr and e-mail: noureddine.zerhouni@ens2m.fr

T. Denœux
Université de Technologie de Compiègne, Heudiasyc, U.M.R. C.N.R.S. 6599, Centre de Recherches de Royallieu, B.P. 20529, F-60205 Compiègne Cedex, France, Address of Institute e-mail: Thierry.Denoeux@hds.utc.fr

sequence or a ball bearing) generating the time-series is a Markov process with unobserved (hidden, latent) discrete states. In HMMs, the states are not visible but when the system is entering in one of the states, the features follow a particular probability distribution. The sequence of observations thus provides information about the sequence of states. One of the most powerful characteristics of HMMs, accounting for its wide range of applications, is the possibility to estimate the parameters efficiently and automatically. Given a training dataset composed of the observed data $\mathbf{X} = \{x_1, \ldots, x_t, \ldots, x_T\}$ (where x_t can be continuous or discrete), and denoting by K the number of hidden states such that the state variable y_t at time t can take a value in

$$\Omega_{\boldsymbol{Y}} = \{1, \ldots, j, \ldots, K\} \ , \tag{1}$$

the following parameters have to be estimated:

- $\boldsymbol{\Phi} = \{\boldsymbol{\phi}_1, \ldots, \boldsymbol{\phi}_j, \ldots, \boldsymbol{\phi}_K\}$ is the set of parameters characterising the probability distribution of observations given each state:

$$b_j(x_t) = P(x_t | y_t = j; \boldsymbol{\phi}_j), j = 1 \ldots K \tag{2}$$

- $\boldsymbol{A} = [a_{ij}]$ with

$$a_{ij} = P(y_t = j | y_{t-1} = i), i = 1 \ldots K, j = 1 \ldots K \tag{3}$$

  that is the probability of the system to be in state j at time-instant t, given the system was in state i at $t-1$, with $\sum_j a_{ij} = 1$.
- $\boldsymbol{\Pi} = \{\pi_1, \ldots, \pi_j, \ldots, \pi_K\}$, where

$$\pi_j = P(y_1 = j) \tag{4}$$

  is the probability of state j at $t = 1$, such that $\sum_i \pi_i = 1$.

In the sequel, all these parameters are aggregated in a vector θ:

$$\theta = \{\mathbf{A}, \boldsymbol{\Pi}, \boldsymbol{\Phi}\} \ . \tag{5}$$

These parameters can be estimated using an iterative procedure called the Baum-Welch algorithm [1, 9] and relying on the Expectation-Maximisation process.

There are applications where some observations x_t in the training data $\mathbf{X}$ are associated to a label that actually represents the state at time t. Instead of considering the labelling process as a binary one, where states can be known or unknown, we address the problem of partially-supervised HMM training, assuming the labels to be represented by belief functions. These functions can represent uncertainty and imprecision about the states and can be in time-series modelling and analysis.

The contribution of this paper holds in the development of a model called Partially-Hidden Markov Model (PHMM) that manages partial labelling of the training dataset in HMMs. Compared to [3], we take into account the temporal dependency into account, helping in time-series modelling. The proposed approach is based on the Evidential Expectation-Maximisation (E2M) algorithm [5].

2 Partially-Hidden Markov Models (PHMM)

Given the observation sequence $\mathbf{X} = \{x_1, x_2, \ldots, x_T\}$, there are three main problems of interest in connection with HMMs [9]:

Problem 1: Given a model $\theta = \{\boldsymbol{\Pi}, \boldsymbol{A}, \boldsymbol{\Phi}\}$, how to compute its likelihood $L(\theta; \mathbf{X})$?
Problem 2: Given a model θ, how to choose the state sequence $\mathbf{Y}^* = \{y_1^*, y_2^*, \ldots, y_T^*\}$ that best explains observations?
Problem 3: How to estimate parameters $\theta = \{\boldsymbol{\Pi}, \boldsymbol{A}, \boldsymbol{\Phi}\}$ of a model?

These problems have been solved in different ways for some decades in HMM [9]. In the sequel, we present the solutions for the case where partial information on states is available in the form of a set of belief functions m defined on the set of states $\Omega_{\boldsymbol{Y}}$. States are then "partially hidden" and the case of completely hidden states is recovered when all the masses are vacuous.

The main idea behind the solutions of partially-supervised training in statistical models is to combine the probability distributions on hidden variables with the belief masses m. This combination can be computed from the contour function pl associated to m.

The next paragraph describes the main features of the E2M algorithm in order to introduce the conditioning process that plays a central role in solutions for problems 1, 2 and 3. The E2M algorithm will be used in the last paragraph dedicated to parameter estimation in PHMMs.

2.1 Generalized likelihood function and E2M algorithm

The Evidential EM (E2M) [5] is an interative algorithm dedicated to maximum likelihood estimation in statistical models based on uncertain observations encoded by belief functions. As for the usual EM algorithm, the E2M algorithm does not maximise directly the observed-data likelihood function denoted here $L(\theta; \mathbf{X}, m)$ but it focuses instead on a lower bound called the auxiliary function [2], and usually denoted by Q and defined as:

$$Q(\theta, \theta^{(q)}) = \mathbb{E}_{\theta^{(q)}} \left[\log L(\mathbf{X}, \mathbf{Y}; \theta) | \mathbf{X}, pl \right] \quad , \tag{6}$$

where pl denotes the contour function associated to m and $\theta^{(q)}$ is the fit of parameter θ at iteration q, and Q represents the conditional expectation of the complete-data log-likelihood. In the E-step of the E2M algorithm, the conditional expectation in the auxiliary function Q is taken with respect to $\gamma' \stackrel{\text{def}}{=} P(\cdot|\mathbf{X}, pl; \theta^{(q)}) = P(\cdot|\mathbf{X}; \theta^{(q)}) \oplus pl$, that is the combination of the expectation, denoted γ_t, with the plausibilities using Demspter's rule [4, 5]. The new expectation is then defined for each state j at time t by $\gamma_t'(j|pl; \theta^{(q)}) = P(y_t = j|\mathbf{X}, pl; \theta^{(q)})$:

$$\gamma_t^{'}(j|pl;\theta^{(q)}) = \frac{\gamma_t(j;\theta^{(q)}) \cdot pl_t(j)}{L(\theta^{(q)};\mathbf{X},pl)} \tag{7}$$

and the auxiliary function becomes:

$$Q(\theta,\theta^{(q)}) = \frac{\sum_{\mathbf{Y}} P(\mathbf{Y}|\mathbf{X};\theta^{(q)}) \cdot pl(\mathbf{Y}) \cdot \log L(\mathbf{X},\mathbf{Y};\theta)}{L(\theta^{(q)};\mathbf{X},pl)} \quad . \tag{8}$$

The M-step is similar to to that of the usual EM algorithm and consists in maximising Q with respect to θ. The maximisation is ensured to increase the likelihood of observed data since E2M inherits the monotonicity of EM as for any sequence $L(\theta^{(q)};\mathbf{X},pl)$, $q = 1,2\ldots$, we have $L(\theta^{(q+1)};\mathbf{X},pl) \geq L(\theta^{(q)};\mathbf{X},pl)$.

2.2 Solution to problem 1 in PHMM

Using a similar process as in usual HMM (see [2] for details on HMM), the marginal posterior distribution on latent variables for the set of parameters $\theta^{(q)}$ at iteration q of E2M can be rewritten as:

$$\gamma_t^{'} = P(y_t|\mathbf{X};\theta^{(q)}) \oplus pl_t = \alpha_t^{'} \cdot \beta_t \tag{9}$$

with $\alpha_t^{'} \stackrel{\text{def}}{=} P(\mathbf{X}_{1:t}, y_t|pl;\theta^{(q)})$ and $\beta_t^{'} \stackrel{\text{def}}{=} P(\mathbf{X}_{t+1:T}|y_t;\theta^{(q)})$. The definition of β remains the same as in the standard algorithm with $\beta_t(i;\theta^{(q)}) = \sum_j \beta_{t+1}(j;\theta^{(q)}) \cdot b_j(x_{t+1}) \cdot a_{ji}, \quad t = 2\ldots T$ starting from $\beta_T(i;\theta^{(q)}) = 1, \forall i$. The probability of jointly observing a sequence $\mathbf{X}_{1:t}$ up to t and state j at time t given the parameters and the uncertain data is given by the modified forward variable $\alpha_t^{'}$ such that $\alpha_t^{'}(j;\theta^{(q)}) = P(\mathbf{X}_{1:t}, y_t = j|pl;\theta^{(q)})$ with:

$$\alpha_t^{'}(j;\theta^{(q)}) = \frac{\alpha_t(j;\theta^{(q)}) \cdot pl_t(j)}{L(\theta^{(q)};\mathbf{X},pl)} \tag{10}$$

and therefore

$$\gamma_t^{'}(j;\theta^{(q)}) = \frac{\alpha_t(j;\theta^{(q)}) \cdot pl_t(j) \cdot \beta_t(j;\theta^{(q)})}{L(\theta^{(q)};\mathbf{X},pl)} \quad . \tag{11}$$

Variables α and β are the same as in HMM [2].

Summing Eq. 11 over latent variables gives the observed data likelihood. Therefore, to assess the likelihood function $L(\theta^{(q)};\mathbf{X},pl)$ at the current iteration of the E2M algorithm, we simply need to choose a time index t. A good candidate is the index T since in this case we do not need to evaluate β_T (that equals to 1) reducing the computation load:

$$L(\theta^{(q)};\mathbf{X},pl) = \sum_{j=1}^{K} \alpha_T(j;\theta^{(q)}) \cdot pl_T(j) \tag{12}$$

Practically, we can use the normalization process proposed in [9] in order to cope with the limited machine precision range.

2.3 Solution to problem 2 in PHMM

The Viterbi algorithm [7] was defined in order to retrieve the best sequence of hidden states within the noisy observations. The best sequence is found in $K^2 \times T$ operations (instead of K^T for a greedy search) and is ensured to be the one with the highest likelihood. Given the observed data $\mathbf{X}$, the Viterbi algorithm finds the maximum a posteriori (MAP) sequence $\mathbf{Y}^* = \{y_1^*, \ldots, y_t^*, \ldots, y_T^*\}, y_t^* \in \Omega_{\mathbf{Y}}$. In PHMM, the MAP criterion is modified by taking soft labels into account, i.e., $P(\mathbf{Y}^*|\mathbf{X}, pl; \theta^{(q)})$ or, equivalently, $\log P(\mathbf{X}, \mathbf{Y}^*|pl; \theta^{(q)})$. In HMMs, the Viterbi algorithm is called the max-sum product algorithm and it is equivalent to a forward propagation with conditioning at each time-step by the potential predecessors of each state. In PHMMs, a similar reasoning can be applied where conditioning (by singletons states) naturally leads to the use of plausibilities. The MAP criterion can be written as:

$$\delta_t^{'}(j; \theta^{(q)}) = \max_i \left[\delta_{t-1}^{'}(i; \theta^{(q)}) \cdot a_{ij}\right] \cdot b_j(x_t) \cdot pl_t(j), \quad t = 2 \ldots T \tag{13}$$

starting from $\delta_1^{'}(j; \theta^{(q)}) = \pi_j \cdot pl_t(j) \cdot b_j(x_1)$. Keeping track of the argument maximising this expression as $\psi_t^{'}(j) = \text{argmax}_i \left[\delta_{t-1}^{'}(i; \theta^{(q)}) \cdot a_{ij}\right]$, the backtracking of the best state sequence ending in $y_t^* = j$ at time t is given by $y_{t-1}^* = \psi_t^{'}(y_t^*)$.

2.4 Solution to problem 3 in PHMM

In the E2M algorithm, the auxiliary function is given by Eq. 8. In order to define the maximisation step, the Q-function has to be computed. For that purpose, we introduce the multinomial representation of variables such that $y_{tj} = 1$ if state j at time t is true, else $y_{tj} = 0$. Then, we can write:

$$P(\mathbf{Y}, \mathbf{X}; \theta) = \left(\prod_{j=1}^{K} \pi_j^{y_{1j}}\right) \cdot \left(\prod_{t=2}^{T}\prod_{i=1}^{K}\prod_{j=1}^{K} a_{ij}^{y_{t-1,i}, y_{tj}}\right) \cdot \left(\prod_{t=1}^{T}\prod_{j=1}^{K} b_j(x_t)^{y_{tj}}\right) \tag{14}$$

Taking the logarithm of the above expression leads to the complete-data log-likelihood. In this paper, partial knowledge on y_{tj} is assumed to be represented by a belief function (and in particular by its contour function $pl_t(j), \forall t = 1 \ldots T, j = 1 \ldots K$). The auxiliary function Q thus becomes:

$$Q(\theta,\theta^{(q)}) = \mathbb{E}_{\theta^{(q)}}\left[\log P(\mathbf{X},\mathbf{Y};\theta)|\mathbf{X},pl\right] \tag{15a}$$

$$= Q_\pi(\theta,\theta^{(q)}) + Q_{\mathbf{A}}(\theta,\theta^{(q)}) + Q_{\boldsymbol{\Phi}}(\theta,\theta^{(q)}) \ , \tag{15b}$$

with $Q_\pi(\theta,\theta^{(q)}) = \sum_{j=1}^{K} \mathbb{E}_{\theta^{(q)}}\left[y_{1j}|\mathbf{X},pl\right]\cdot\log\pi_j$ given by:

$$Q_\pi(\theta,\theta^{(q)}) = \sum_{j=1}^{K} \gamma_1^{'}(j;\theta^{(q)})\cdot\log\pi_j \ , \tag{16}$$

and $Q_{\mathbf{A}}(\theta,\theta^{(q)}) = \sum_{t=2}^{T}\sum_{i=1}^{K}\sum_{j=1}^{K}\mathbb{E}_{\theta^{(q)}}\left[y_{t-1,i}y_{tj}|\mathbf{X},pl\right]\cdot\log a_{ij}$ with:

$$Q_{\mathbf{A}}(\theta,\theta^{(q)}) = \sum_{t=2}^{T}\sum_{i=1}^{K}\sum_{j=1}^{K}\xi_{t-1,t}^{'}(i,j;\theta^{(q)})\log a_{ij} \ , \tag{17}$$

and $Q_{\boldsymbol{\Phi}}(\theta,\theta^{(q)}) = \sum_{t=1}^{T}\sum_{j=1}^{K}\mathbb{E}_{\theta^{(q)}}\left[y_{tj}|\mathbf{X},pl\right]\cdot\log b_j(x_t)$ given by:

$$Q_{\boldsymbol{\Phi}}(\theta,\theta^{(q)}) = \sum_{t=1}^{T}\sum_{j=1}^{K}\gamma_t^{'}(j;\theta^{(q)})\cdot\log b_j(x_t) \ . \tag{18}$$

In the above expressions we have:

$$\gamma_t^{'}(j;\theta^{(q)}) = \frac{\gamma_t(j;\theta^{(q)})\cdot pl_t(j)}{\sum_{l=1}^{K}\gamma_t(l;\theta^{(q)})\cdot pl_t(l)} \ , \tag{19}$$

which is the marginal posterior distribution of a latent variable y_j at t given pl, and

$$\xi_{t-1,t}^{'}(i,j;\theta^{(q)}) = \frac{\xi_{t-1,t}(i,j;\theta^{(q)})\cdot pl_{t-1}(i)\cdot pl_t(j)}{\sum_{l=1}^{K}\xi_{t-1,t}(i,l;\theta^{(q)})\cdot pl_{t-1}(i)\cdot pl_t(l)} \tag{20}$$

is the joint probability of two consecutive latent variables $y_{t-1,i}$ and y_{tj} given pl. The optimal parameters at each iteration of E2M are given by using a similar reasoning as in the standard algorithm, but the posterior probability over latent variables now depends on the plausibilities:

$$\pi_j^{(q+1)} = \frac{\gamma_1(j;\theta^{(q)})\cdot pl_1(j)}{\sum_{l=1}^{K}\gamma_1(l;\theta^{(q)})\cdot pl_1(l)} \tag{21a}$$

$$a_{ij}^{(q+1)} = \frac{\sum_{t=2}^{T}\xi_{t-1,t}(i,j;\theta^{(q)})\cdot pl_{t-1}(i)\cdot pl_t(j)}{\sum_{t=2}^{T}\sum_{l=1}^{K}\xi_{t-1,t}(i,l;\theta^{(q)})\cdot pl_{t-1}(i)\cdot pl_t(l)} \tag{21b}$$

The maximisation of $Q_{\boldsymbol{\Phi}}(\theta,\theta^{(q)})$ depends on the form of the distribution of observations given the latent variable j.

3 Partial results, conclusion and further work

Partial results: To illustrate this approach, we considered that observations can be modelled by mixtures of Gaussians. We proceeded as in standard HMM to derive the M-step in PHMM and to estimate the parameters of the distributions. Equations are however not reported in this paper.

For illustration purpose, we used the dataset of the PHM'08 data challenge [12] concerning the health state of a turbofan engine. It was manually segmented into four states (to evaluate the results) such that each time-series is accompanied by a set of labels reflecting the current state of the fan, that is normal, transition, degrading or faulty mode. Each label corresponds to a mass function focused on a singleton, except in the transitions where doubt between two labels is defined. The segmentation and the associated BBA are available at `http://www.femto-st.fr/~emmanuel.ramasso/PEPS_INSIS_2011_PHM_by_belief_functions.html`. The BBA were then transformed into plausibilities. For these tests, we corrupted them by additive noise: $pl_t(j) \leftarrow pl_t(j) + \sigma_k \cdot \varepsilon_t(j)$, where $\sigma_k \in \{0, 0.1, \ldots, 1\}$ and $\varepsilon_t(j) \sim \mathscr{U}_{[0,1]}$ was drawn from a uniform distribution. For each noise level, we considered the influence of the number of unlabelled data $\nu_k \in \{0\%, 10\%, \ldots, 100\%\}$. The partitioning of time-series in the testing dataset estimated by HMM and PHMM using the Viterbi algorithm as defined in HMM (since we do not know the labels for the testing) were compared using the Folkes and Mallows index ($F \in [0,1]$) [11] by computed the relative performance improvement $G = F_{pshmm}/F_{hmm} - 1$ with $G \in [-1,1]$ such that if $G > 0$ (resp. $G < 0$), the proposed PHMM provided a better (resp. worse) segmentation of the time-series into states.

The evolution of G is given in Figure 1 that shows an improvement by several percents when using the proposed PHMM (up to 12%). When all data were unlabelled and with no noise (bottom right hand-side corner), both models provided exactly the same results, as expected. When the noise increased, the performance decreased but was still higher than that of the standard HMM. The most difficult cases were encountered when the noise was high (top of figure), where PHMM improvements were between $[2\%, 5\%]$.

Conclusion and further work: Taking partial knowledge into account is of crucial importance in many statistical models. Encoding prior information by belief functions leads to simple modifications of the initial estimation formula while remaining theoretically sound. The statistical model considered in this paper was the Hidden Markov Models. Further work remains to be done in order to compute in developing reestimation formula for various distributions of observations given latent states.

Acknowledgements: This work was partially supported by a PEPS-INSIS-2011 grant from the French National Center for Scientific Research (CNRS) under the administrative authority of the French Ministry of Research.

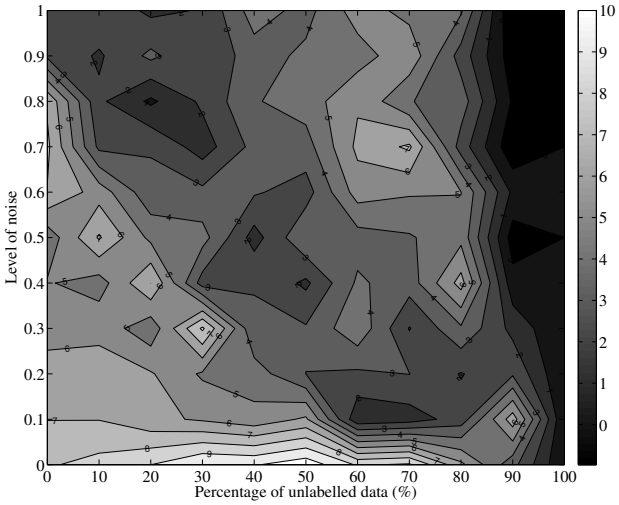

Fig. 1 Performance (G-index): median value over 10 runs with different initialisation. Positive value reflects an improvement provided by PHMM. Here almost all values are positive except darkest areas.

References

1. Baum, L.E., Petrie, T., Soules, G., Weiss, N.: A maximization technique occurring in statistical analysis of probabilistic functions of markov chains. Ann. Math. Stat. **41**, 164–171 (1970)
2. Bishop, C.: Pattern Recognition and Machine Learning. Springer (2006)
3. Côme, E., Oukhellou, L., Denoeux, T., Aknin, P.: Learning from partially supervised data using mixture models and belief functions. Pattern Recognition **42**, 334–348 (2009)
4. Dempster, A.: Upper and lower probabilities induced by multiple valued mappings. Annals of Mathematical Statistics **38**, 325–339 (1967)
5. Denoeux, T.: Maximum likelihood estimation from uncertain data in the belief function framework. IEEE Tr. on Knowledge and Data Engineering (2011). DOI 10.1109/TKDE.2011.201
6. Dong, M., He, D.: A segmental hidden semi-markov model (hsmm)-based diagnostics and prognostics framework and methodology. Mechanical Systems and Signal Processing **21**, 2248–2266 (2007)
7. Forney, G.: The viterbi algorithm. Proceedings of the IEEE **61**(3), 268–278 (1973)
8. Murphy, K.P.: Dynamic Bayesian networks: Representation, inference and learning. Ph.D. thesis, UC Berkeley (2002)
9. Rabiner, L.: A tutorial on hidden Markov models and selected applications in speech recognition. Proc. of the IEEE **77**, 257–285 (1989)
10. Ramasso, E.: Contribution of belief functions to hidden markov models. In: IEEE Workshop on Machine Learning and Signal Processing, pp. 1–6. Grenoble, France (2009)
11. Saporta, G., Youness, G.: Comparing two partitions: Some proposals and experiments. In: COMPSTAT (2002)
12. Saxena, A., Goebel, K., Simon, D., Eklund, N.: Damage propagation modeling for aircraft engine run-to-failure simulation. In: Int. Conf. on Prognostics and Health Management, pp. 1–9. Denver, CO, USA (2008)